2025 年【科學探究競賽-這樣教我就懂】

大專/社會組　科學文章格式

文章題目：做研究別只看 *p* 值！論效果量的重要性

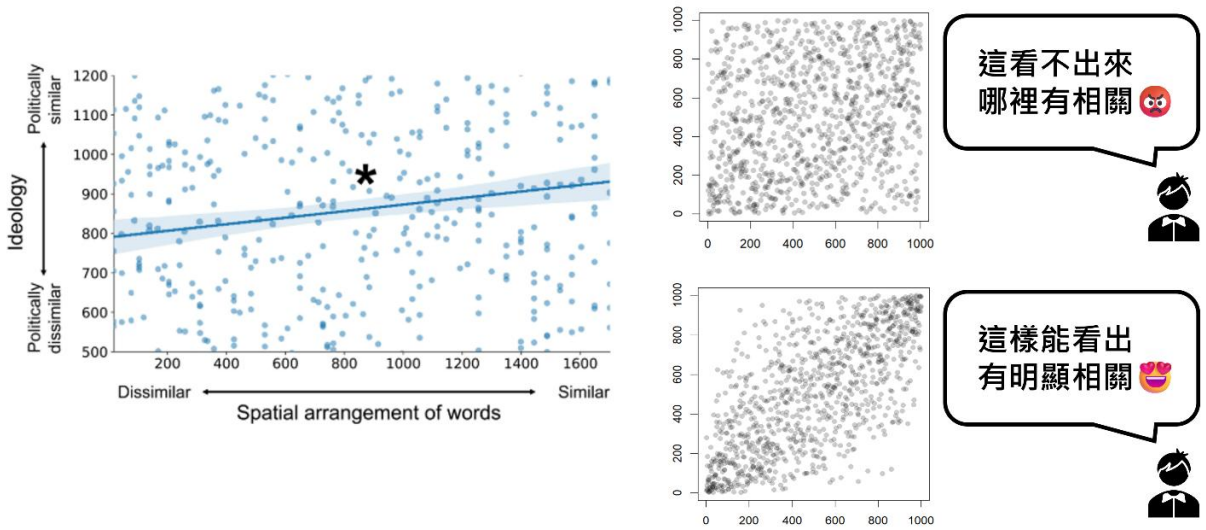
摘要：假設檢定中，樣本數變大時，*p* 值會變得容易顯著。效果量更能反映變項之間的實質關聯強度。

文章內容：（限 500 字~1,500 字）

不論是在教學、研究、或實務應用中，我們常常會使用推論統計方法來看看不同情況下的群體表現是不是有差異。像是不同性格的人，在小組報告貢獻是否一樣多？不同行業工作的人，在挑選伴侶時的偏好特徵是否有所差別？這時候會用到的方法，可能是比兩組人的 *t* 檢定、比多組人的變異數分析，或者是尋找變項之間關聯性的相關分析等。而這些方法，都牽涉到使用假設檢定來提供問題的答案。

但是，不論是使用哪種統計方法，大多數人在解讀統計檢定結果時，只會盯著「*p* 值」（*p*-value）這個數字看——也就是「偽陽性的機率」，通常以「*p* 值小於 5%」（$p < 0.05$）作為判斷標準。當看到 *p* 值過了這個門檻，很多人就會說：「這有顯著差異」。為了讓讀者方便閱讀統計結果，在 $p < 0.05$ 時，有些人甚至會在圖表上加上米字記號（*）來強調。可是，統計上的「顯著」只代表在虛無假設下偶然發生的機率很低，並不等於結果在實務上具有「實質重要性」。也就是說，只關注 *p* 值的做法，其實太草率了！

舉個例子，Bruin 等人（2023）的腦神經研究探討不同政治立場的人，對於同溫層和不同立場的內容反應，有沒有行為上或大腦上的差別。其中一項探討變項相關的檢定統計量 $t = 2.66$，$p = 0.006$，具有顯著關聯性。然而，如果從資料散布圖（見圖一左）來看，就會發現其實資料點非常分散，用肉眼看不出來差別。但因為樣本數非常大（超過 500 筆），即使差異很小，檢定統計量仍然會被放大，檢定結果也會變得容易顯著。實際上，這筆資料的相關係數（*r*）只有 0.11，屬於弱相關，基本上沒有太大實質意義。



圖一　Bruin 等人（2023）研究結果之散布圖（左圖）與作者模擬資料之散布圖（右圖）

我自己也模擬了資料(見圖一右上)，設定兩個變項之間存在比較低的關聯性($r = 0.22$)，並將樣本數設定成 1000 筆。當我對兩個變項的關聯性進行假設檢定時，結果是 $t = 7.04$，$p < 0.001$，看起來非常顯著！然而，我們都知道，這只是因為樣本數夠大，不是真的有很強的關聯。我們真正想要看到的，應該是像圖一右下所呈現的，肉眼就可以判斷兩變項之間存在明顯的共變趨勢（$r = 0.71$）。

這些例子說明了一件事：光看 p 值絕對是不夠的！它可能會讓我們誤以為某個結果很重要，但實際上只是樣本數太多，變得容易顯著而已。所以，我們需要更可靠的指標，那就是——「效果量」（effect size）。效果量可以告訴我們，一個關係或差異「有多大」，而不是只問「存不存在」關聯或差異。它不像 p 值那樣會受到樣本數太大的影響，因此更能反應資料蘊含的實際意義。而從假設檢定所延伸出來的效果量公式，都有包含考量樣本數的參數，這有助於減弱太多筆資料所帶來的影響。

那麼，要怎麼看效果量？如果是相關係數，一般認為 0.1 是小效果，0.3 是中效果，0.5 以上才算大效果（Cohen, 1988）。但這些只是一種參考標準，這些效果量的強度定義，是基於統計學資深學者們的主觀經驗。要如何決定哪種效果視為不錯的實質意義，學界也並沒有統一客觀的標準，實際上要看研究背景來作判斷。根據我自己的經驗，效果量如果沒有達到中等強度（如：$r < 0.3$），就不值得太在意。

順帶一提，如果你對各種統計檢定的效果量計算與解讀有興趣，可以看看 Jané 等人（2024）整理的網站（各種效果量的強度標準，見圖二），上面有很多詳細且清楚的教學與公式唷！

Effect Size	Reference	Small	Medium	Large
Mean Differences				
Cohen's d or Hedges' g	Cohen (1988)[1]	0.20	0.50	0.80
		0.18	0.37	0.60
	Lovakov and Agadullina (2021)[2]	0.15	0.36	0.65
Correlational				
Correlation Coefficient (r)	Cohen (1988)	.10	.30	.50
	Richard, Bond Jr., and Stokes-Zoota (2003)[3][4]	.10	.20	.30
	Lovakov and Agadullina (2021)	.12	.24	.41
	Paterson et al. (2016)	.12	.20	.31
	Bosco et al. (2015)	.09	.18	.26
Cohen's f^2		.02	.25	.40
eta-squared (η^2)	Cohen (1988)	.01	.06	.14
Cohen's f	Cohen (1988)	.10	.25	.40
Categorical				
Cohen's w	Cohen (1988)	0.10	0.30	0.50
Phi	Cohen (1988)	.10	.30	.50
Cramer's V		[5]		
Cohen's h	Cohen (1988)	0.2	0.5	0.8

圖二　Jané 等人（2024）統整的檢定效果量之強度標準

參考資料

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Academic Press. https://doi.org/10.4324/9780203771587

de Bruin, D., van Baar, J. M., Rodríguez, P. L., & FeldmanHall, O. (2023). Shared neural representations and temporal segmentation of political content predict ideological similarity. *Science Advances*, *9*(5), eabq5920. https://doi.org/10.1126/sciadv.abq5920

Jané, M., Xiao, Q., Yeung, S., Ben-Shachar, M. S., Caldwell, A., Cousineau, D., Dunleavy, D. J., Elsherif, M., Johnson, B., Moreau, D., Riesthuis, P., Röseler, L., Steele, J., Vieira, F., Zloteanu, M., & Feldman, G. (2024). Guide to Effect Sizes and Confidence Intervals. http://dx.doi.org/10.17605/OSF.IO/D8C4G

註：

1. 未使用本競賽官網提供「科學文章表單」格式投稿，**將不予審查**。
2. 字數沒按照本競賽官網規定之限 500 字~1,500 字，**將不予審查。**
   PS.摘要、參考資料與圖表說明文字不計入。
3. 建議格式如下：
   - 中文字型：微軟正黑體；英文、阿拉伯數字字型：Times New Roman
   - 字體：12pt 為原則，若有需要，圖、表及附錄內的文字、數字得略小於 12pt，不得低於 10pt
   - 字體行距，以固定行高 20 點為原則
   - 表標題的排列方式為向表上方置中、對齊該表。圖標題的排列方式為向圖下方置中、對齊該圖